Global Environmental Change 49 (2018) 186–196

Contents lists available at ScienceDirect

Global Environmental Change

journal homepage: www.elsevier.com/locate/gloenvcha

What makes for compelling science? Evidential diversity in the evaluation of scientific arguments☆

Arthur Kary*, Ben R. Newell, Brett K. Hayes

University of New South Wales, Sydney, Australia

ARTICLE INFO

Keywords:
Reasoning
Science communication
Diversity effect
Best-worst scores
Discrete choice experiment
Climate change
Public health

ABSTRACT

Despite overwhelming scientific evidence, many members of the public remain skeptical about anthropogenic global warming. Hence, we examined how the presentation of factual scientific evidence affects lay evaluations of scientific claims. Taking inspiration from cognitive research on inductive reasoning, two studies examined the impact of evidential diversity on acceptance of claims in the domains of climate change and public health. Participants were presented with scientific claims based on competing evidence options and were asked to choose the best and worst form of evidence for each claim. The diversity of the available evidence was manipulated across three dimensions; geographical (evidence from two geographically near or far nations), socio-cultural (evidence from two culturally similar or dissimilar nations), and temporal (evidence drawn from two different periods or the same period). In both studies, diverse evidence on the geographical and socio-cultural dimension increased perceived support for scientific claims, but the relative impact of these dimensions differed between domains; geographical diversity had a larger effect on claims about climate change; socio-cultural diversity had a larger effect on claims about health. On the temporal dimension, recent non-diverse evidence (i.e. from the same recent period) increased perceived support for scientific claims more than diverse evidence. These results may have important implications for the communication of complex scientific evidence to a lay audience.

1. Introduction

The relevance of science to daily life and everyday decisions is growing. With increasing frequency, lay people are confronted with making decisions about complex socio-scientific issues such as whether they will support government regulation of carbon emissions, whether they should install solar paneling in their home, or whether they should reduce the amount of salt in their diet (Sinatra and Hofer, 2016).

Despite the increasing prevalence of socio-scientific issues, in many areas there remains a significant gap between the scientific and public understanding of key problems. A signature case is belief in anthropogenic global warming (AGW). Over the past decade there has been increasing agreement among scientists that carbon emissions from human activity contribute to global warming (Freudenburg and Muselli, 2010; IPCC, 2014). Over the same period however, public belief in AGW in countries such as the USA and Australia has remained static or even decreased (Brulle et al., 2012; Leviston et al., 2015; Smith and Leiserowitz, 2012).

Such cases highlight the importance of understanding how non-scientists interpret and evaluate scientific evidence. Despite the crucial importance of this issue, surprisingly little work has investigated the cognitive factors that affect how non-scientists evaluate the strength of scientific arguments (Corner and Hahn, 2009; Hahn et al., 2016). For example, although provision of scientifically accurate information about AGW can increase acceptance of the phenomenon and promote pro-environmental attitudes (Ranney and Clark, 2016), there have been few attempts to examine the cognitive mechanisms involved in the lay interpretation of such information. An understanding of these mechanisms is important if we wish to maximize the acceptance and impact of science in public policy debates.

A potentially useful approach to advancing our understanding of how lay people evaluate scientific data involves considering the implications of research on inductive reasoning in non-scientific domains. How people use evidence to draw more general conclusions is the central focus of research on inductive reasoning (Hayes et al., 2010). In studies of inductive reasoning, people are presented with new evidence that is assumed to be true (e.g., that lions and dolphins have “beta cells inside”) and are asked to evaluate the strength or plausibility of some conclusion based on this evidence (e.g., that all mammals have beta cells). Such inferences are probabilistic, with belief in the conclusions

☆ This work was supported by Australian Research Council Linkage Grant LP120100224 to the second and third authors. We thank Jeremy Ngo and Manfred Ng for their help in manuscript preparation and Marilynn Brewer and Rachel McDonald for useful discussions during the course of this project.

* Corresponding author at: School of Psychology, University of New South Wales, Sydney 2052, Australia.
E-mail address: a.kary@unsw.edu.au (A. Kary).

https://doi.org/10.1016/j.gloenvcha.2018.01.004
Received 25 January 2017; Received in revised form 3 January 2018; Accepted 28 January 2018
Available online 14 March 2018
0959-3780/

increasing or decreasing depending on the quality and quantity of evidence.

Laboratory studies of inductive reasoning have revealed a range of factors that affect perceptions of argument strength (see Hayes and Heit, 2013 for a review). In this respect, the literature on inductive reasoning is a potentially rich source of ideas about what types of evidence people find more or less convincing. However, in order to minimize the impact of prior knowledge on argument evaluations, the "evidence" and conclusions presented in these laboratory studies are usually not factual, often relying on invented or abstract properties (e.g., "has beta cells" or "has property P"). A major aim of the current studies was to examine whether a key factor identified in laboratory work as influencing inductive inferences, *evidence diversity*, also affects the way that people evaluate factual scientific evidence.

1.1. Evidence diversity in inductive reasoning

Philosophers of science have argued that other things being equal, a scientific theory is more strongly supported by diverse rather than non-diverse evidence (Bacon, 1620/1898; Hempel, 1966; Salmon, 1984; see Heit et al., 2004 for a review). Salmon (1984) for example, reviews the way that early 20th century scientists developed a range of methods for deriving Avogadro's number (6.02×10^{23}), the number of particles in a mole of any substance. These included Brownian movement, alpha particle decay, X-ray diffraction, black body radiation, and electrochemistry. The derivation of Avogadro's number was a significant scientific discovery as it provided support for what was, at the time, a controversial hypothesis: the existence of atoms and molecules. Salmon argued that the evidence taken from any one experimental technique alone would be unlikely to be viewed as convincing evidence in favor of atomic theory. The convergence of results based on diverse methods greatly strengthened the theory.

Research on inductive reasoning shows that non-scientists view evidential diversity as important in evaluating the strength or plausibility of an argument. Although there are some interesting exceptions (e.g., Medin et al., 2003), the general finding is that conclusions based on diverse forms of evidence are viewed as more plausible or convincing than those based on more similar forms of evidence (e.g., Feeney and Heit, 2011; Kim and Keil, 2003; Osherson et al., 1990; Shafto et al., 2007). For example, people are more likely to endorse a general conclusion (e.g., that mammals have some property P) given evidence about diverse or dissimilar instances (e.g., lions and dolphins have property P) than evidence about very similar instances (e.g., lions and tigers have property P). Such sensitivity to evidential diversity emerges during the elementary school years (Heit and Hahn, 2001; Rhodes et al., 2010) and has been shown to affect evidence selection in hypothesis testing (López, 1995), memory for learned material (Hahn et al., 2005), and conceptual change during development (Hayes et al., 2003). Although there is some debate about whether such effects are normative (cf. Heit and Hahn, 2001; Lo et al., 2002; Wayne, 1995), there seems little doubt that evidence diversity is an influential heuristic for assessing the strength of an argument.

These findings lead to the prediction that factual scientific arguments supported by more diverse forms of evidence will often be viewed as more plausible and persuasive than arguments based on less diverse forms of evidence. Here we test this prediction in two experiments using arguments from two scientific domains: climate change and public health.

1.2. Assessing the diversity of complex scientific evidence

The question of what constitutes "diverse evidence" is almost certainly more complex in real scientific domains than has been the case in previous laboratory studies of inductive reasoning. In previous work, the problem of measuring evidential diversity has generally been addressed by ensuring that all argument premises and conclusions are drawn from a single conceptual domain (e.g., animals). Hence, instances can be compared on a single dimension (e.g., taxonomic similarity) and it is quite straightforward to establish the relative similarity or diversity of the instances used in argument premises.

In contrast, the evidence involved in scientific domains such as climate change and public health is notoriously complex. Climatic events and population-wide changes in health are the end result of the interaction between multiple causal systems. The scientific evidence that bears witness to these phenomena comes in a wide variety of forms including archival records, observations from geographically and socially distinct testing sites, and experimental findings.

The comparison of two or more types of evidence relating to actual scientific phenomena therefore is likely to involve consideration of multiple dimensions of similarity. For example, suppose we learned the following: a) over the first half of the 20th century there was an increase in the mean sea level around the Australian coastline, and b) records taken over the past 20 years have shown that mean sea level has increased along the coastline of West Africa. Unlike the evidence used in conventional induction research, these statements differ on at least three dimensions; temporal (data collected at different times), geographical (data collected from different parts of the world), and social (data collected from regions with different cultural and economic profiles). Each of these dimensions needs to be considered in assessing the diversity of this evidence and how it affects belief in a conclusion like "global sea levels are rising".

This example raises the question of how people respond to diversity on multiple dimensions. Extrapolating from previous studies of the diversity heuristic in arguments with a single conceptual dimension, one might predict that diverse evidence across multiple dimensions will have a cumulative effect on confidence in a scientific claim. In other words, diverse evidence across multiple dimensions provides more compelling evidence for a claim than diverse evidence on fewer dimensions.

An alternative possibility is that in complex scientific domains people are selective in the way they employ the diversity heuristic. Diversity on some dimensions may have a greater impact on the evaluation of scientific claims than others. Nonscientists understand that different causal factors operate in different scientific domains (Sloman et al., 2007). For example, an understanding of the different causal principles that operate in the domains of everyday physics, biology and psychology emerges relatively early in development (Carey, 1995; Keil, 2003). Hence, the dimensions of diverse evidence that are perceived as most relevant for evaluating a scientific conclusion are likely to vary across domains. In the above example, it may be that people are more persuaded to believe that global sea levels are rising by the fact that increases have been detected in two very different locations and over different measurement periods. The social and cultural diversity of the testing sites may be seen as less crucial. In contrast, if the evidence and associated conclusions were concerned with a public health issue (e.g., disease prevalence, rates of immunization) then socio-cultural diversity may be a central consideration.

To take account of these complexities we developed a novel approach to studying evidential diversity in scientific domains. This involved systematically manipulating the diversity of scientific facts on three dimensions: temporal, geographical and social. Diversity on these dimensions was manipulated factorially so that people saw evidence items that were diverse on zero, one, two or all three dimensions. This allowed us to examine the separate impact of diversity on each dimension on belief in scientific conclusions in the domains of climate change and public health.

An additional justification for focusing on these three dimensions was that they are among the key factors that have been shown to influence subjective perceptions of "psychological distance" (see McDonald et al., 2015 for a review). Psychological distance refers to the extent to which something is perceived as distant from oneself. Previous work shows that the perceived "distance" of events (e.g., climate

change, disease epidemics) influences beliefs about the personal significance of the phenomena and the perceived need for ameliorative action (Fujita et al., 2014; Weber, 2010). Uncovering the most salient dimensions of evidential diversity relating to climate change and public health therefore has the additional benefit of indicating the types of scientific evidence that are most likely to promote concern and action in each domain.

2. Experiment 1

In studies of diversity-based reasoning, argument evaluation has typically been measured by binary choices between diverse and non-diverse arguments (e.g., López, 1995) and/or ratings of the strength of arguments containing diverse or non-diverse evidence (e.g., Feeney and Heit, 2011). The use of such methods can be problematic however when evaluating judgments about arguments that differ on multiple dimensions. For each binary dimension included in the design the number of arguments per conclusion increases exponentially, meaning it is not always practical to devise pairwise choices for all combinations of arguments. Moreover, in reasoning experiments it is desirable to examine the diversity effect for more than one conclusion, so the number of choices presented to participants can become unwieldy very quickly.

To avoid such problems we employed a best-worst choice paradigm (Louviere et al., 2015). The best-worst choice paradigm is a type of discrete choice experiment, in which participants' preferences among options are measured by a series of choices between subsets of those options. Modern theoretical and methodological developments have allowed for the estimation of these preferences without needing to present all possible choices to participants (e.g., Street et al., 2005). The paradigm has been refined such that estimates of participant preferences can be reliably elicited from a manageable number of choice trials for each participant. This balance of reliability and expediency makes it very attractive for our purposes (Louviere et al., 2008).

In the current paradigm, three evidence options that varied in geographical, social and temporal diversity were presented. Participants were asked to choose which of the three options provided the most compelling ("best") evidence in favor of a scientific claim (e.g., "human activity is contributing to the accumulation of CO_2 in the atmosphere") and which option provided the least compelling ("worst") evidence. Comparison of best-worst ratings between evidence options allowed us to evaluate the impact of geographical, social and temporal diversity on endorsement of scientific claims in the health and climate domains relative to non-diversity. The method also allows us to compare how much the different dimensions of diversity influenced choices. Best-worst choice designs have been successfully employed to uncover the dimensions that guide choice preferences in domains such as marketing (e.g., Cohen, 2009) and healthcare (e.g., Jones et al., 2015).[1]

We predicted that evidence diversity would affect perceptions of the strength of scientific arguments but that this effect would be selective across dimensions of diversity and domains. Evidence on those domains that were most likely to be seen as causally relevant to a given domain (e.g., geographical diversity in climate change; social diversity in public health) were expected to have the greatest impact on evaluation of scientific claims. We were less certain about the impact of temporal diversity on each domain. However, we note that in the climate domain, evidence concerning continuity and change in relevant metrics (e.g. global temperatures, sea levels) over long time scales is often frequently cited in support of claims about current and future trends in global warming (e.g., Mann et al., 1998; Sutton et al., 2015).

2.1. Method

2.1.1. Participants

63 Australian first year Psychology undergraduates participated for course credit. The mean age of the sample was 19.2 (SD = 2.02) and there were 44 females.

2.1.2. Materials

On each trial participants were presented with an argument conclusion (described as a "scientific claim") and three alternative evidence options that could be seen as supporting the claim. Each option contained statements containing information that varied in diversity on three dimensions: temporal, geographical and social. Temporal diversity was manipulated by citing two samples of evidence from either the same, recent time-period (e.g., 2001–present) or different time-periods (e.g., 1950–2000; 2001–present). Geographical diversity was manipulated by citing evidence from geographically close (e.g., Australia and New Zealand) or distant locations (e.g., Australia and Saudi Arabia). Social diversity was manipulated by citing evidence from countries that were relatively similar from a socio-cultural perspective (e.g., Australia and England) or dissimilar (e.g., Australia and Papua New Guinea). For each pair, there was always one premise which related to Australia that did not change across the arguments (we refer to this below as the fixed premise of the arguments). This was because the experiment was conducted in Australia, and thus we defined social and geographical diversity relative to Australia.[2] Table 1 gives examples of premise pairs and conclusions from the climate and health domains. As far as was practicable, the evidence described was factual and was drawn from a variety of on-line sources relevant to either climate change (e.g., Commonwealth Scientific and Industrial Research Organization) or public health (e.g., United Nations, Department of Economic and Social Affairs, Population Division).[3]

The alternative evidence options were generated according to the structure shown in Table 2. As shown in the table, eight sets of evidence options were constructed to allow us to test the effects of evidence diversity on each dimension on perceived support for the scientific claim. We used a simple model-free descriptive statistic known as best minus worst scores (B-W scores) for our analysis (cf. Louviere et al., 2015; Marley et al., 2016). For example, the B-W score for geographical diversity was the number of times evidence options featuring geographical diversity (e.g., option 3 in sets 1 and 2 from Table 2) were chosen as best minus the number of times they were chosen as worst, divided by the number of opportunities to select a geographically diverse option from any of the sets − in this case 8. Thus, the B-W score was a number between −1 and +1, with positive numbers indicating preference for conclusions based on diverse evidence, negative numbers indicating preference for conclusions based on non-diverse evidence, and numbers close to 0 indicating that diversity did not influence choices. This approach utilized a choice structure developed by Street et al. (2005) for discrete choice experiments, and allowed us to test for main effects of the different dimensions of diversity in our best-worst design.

Within each of the climate and health domains, there were four target conclusions (see Table 1 for examples and Appendix A for all conclusions). Items were generated by applying the evidence option structure shown in Table 2 to each of these conclusions, giving rise to a

[1] Note that the best/worst labels are relative and only refer to the strongest and weakest forms of evidence amongst the given options. Evidence that is identified as "worst" may still be seen as supporting a scientific claim.

[2] We conducted a post-hoc manipulation check (after the completion of both experiments) to ensure that our categorization of countries as socially and geographically diverse was consistent with the beliefs of our student population. We found that 24 participants who did not participate in the original experiments unanimously agreed with our classifications. See Supplementary materials for survey method.

[3] See Supplementary materials for a list of example sources. Note that in some cases minor adjustments were made to reduce numerical differences between facts that were paired. The main concern here was to minimize the impact on ratings of evidential strength of differences in the numerical magnitude of change in the measurements described in an argument.

Table 1
Examples of target conclusions ('scientific claim') and evidence options in each domain. For each option the level of diversity (0 or 1) on the respective temporal, geographical and social dimensions is given. For example, Option 2 "1 0 1" contains diverse evidence on the temporal dimension ('1' – evidence from two different time periods; 1950–2000 and post 2001) and social dimensions ('1' evidence from two culturally different nations; Australia and Indonesia), but non-diverse evidence on the geographical dimension ('0' – evidence from two geographically proximate nations; Australia and Indonesia). (See Supplementary materials for data supporting assumptions about the social and geographic proximity of the different countries used in the options.).

Domain	Target Conclusion	Option 1	Option 2	Option 3
Climate	The global climate is warming	0 0 0 • In Australia, each decade from 1950 to 2000 was warmer than the one before • From 1950 to 2000, New Zealand temperatures increased by 0.5C	1 0 1 • In Australia, each decade from 1950 to 2000 was warmer than the one before. • Temperatures in Indonesia have risen 0.14 °C since 2001	1 1 0 • In Australia, each decade from 1950 to 2000 was warmer than the one before. • Since 2001, 6 out of 10 of England's warmest years on record have occurred.
Climate	The global climate is warming	0 0 1 • In Australia, each decade from 1950 to 2000 was warmer than the one before. • Average maximum temperatures in Papua New Guinea increased by 0.11 °C per decade from 1950 to 2000.	1 0 0 • In Australia, each decade from 1950 to 2000 was warmer than the one before. • The mean temperature of New Zealand since 2001 has been 0.26 °C above the historical average	1 1 1 • In Australia, each decade from 1950 to 2000 was warmer than the one before. • Since 2001, Saudi Arabia has broken a heat record at least once.
Public health	The incidence of diabetes is increasing across the globe	0 0 0 • Since 1990, the incidence of diabetes in Australia has increased from 1.5% of the population to 4.2%. • The number of New Zealanders with diabetes has increased from 86 000 in 1990 to 210 000	1 0 1 • Since 1990, the incidence of diabetes in Australia has increased from 1.5% of the population to 4.2%. • From 1980-89, the proportion of Papua New Guineans with diabetes rose from about 8% to about 9.4%	1 1 0 • Since 1990, the incidence of diabetes in Australia has increased from 1.5% of the population to 4.2%. • The proportion of Americans with diabetes rose from about 5.6% in 1980 to about 7.1% by 1989
Public health	The incidence of diabetes is increasing across the globe	0 0 1 • Since 1990, the incidence of diabetes in Australia has increased from 1.5% of the population to 4.2%. • In Fiji, the incidence of diabetes has risen from less than 16% in 1990 to 40%	1 0 0 • Since 1990, the incidence of diabetes in Australia has increased from 1.5% of the population to 4.2%. • The proportion of New Zealand women with diabetes rose from 5.6% to 5.7% from 1980 to 89.	1 1 1 • Since 1990, the incidence of diabetes in Australia has increased from 1.5% of the population to 4.2%. • From 1980-89, the proportion of Cubans with diabetes rose from about 8.4% to about 8.9%

Table 2
Summary of the structure of evidence options. For each option the level of diversity (0 = non-diverse; 1 = diverse) on each of the temporal, geographical and social dimensions is given. Contrasts taken from Street et al. (2005).

Set	Option 1			Option 2			Option 3		
	Time	Geo	Social	Time	Geo	Social	Time	Geo	Social
1	[0	0	0]	[1	0	1]	[1	1	0]
2	[0	0	1]	[1	0	0]	[1	1	1]
3	[0	1	0]	[1	1	1]	[1	0	0]
4	[0	1	1]	[1	1	0]	[1	0	1]
5	[1	0	0]	[0	0	1]	[0	1	0]
6	[1	0	1]	[0	0	0]	[0	1	1]
7	[1	1	0]	[0	1	1]	[0	0	0]
8	[1	1	1]	[0	1	0]	[0	0	1]

total of 32 items in each domain (see supplementary materials for all items including target conclusions and evidence options).

2.1.3. Procedure

On each trial participants were presented with an argument which consisted of a) a scientific claim (the argument conclusion) at the top of a computer screen and b) three evidence options below (see Fig. 1). Each option consisted of two statements that contained diverse or non-diverse temporal, geographical and social evidence supporting the scientific claim (see Tables 1 and 2), with the order of these statements randomized across trials. Participants were asked to choose which of the three options provided the "best" evidence for the claim and which of the options provided the "worst" evidence for the claim. They could not select the same option for each question.

Eight such trials were presented in a block for each scientific claim, with each trial using one of the evidence sets shown in Table 2. Trial order within these blocks was randomized. There were four scientific claims that were evaluated for each of the climate and public health domains, so all participants completed 64 trials. Block order was randomized so on adjacent blocks participants often evaluated claims relevant to different domains.

In addition to evaluating evidence based on diversity, it was also important to check that the various target claims or conclusions were viewed as equally plausible in the absence of any evidence. To this end participants were presented with the 8 target scientific claims in random order, and indicated how likely they thought the statement was to be true on a scale of 0 (very unlikely to be true) to 100 (very likely to be true). We counterbalanced the order of the tasks, such that half of the participants completed these ratings before the best-worst diversity task and half afterwards. We gave no explicit instruction to participants that there would be two tasks relating to the same conclusions, so we expected that participants would treat the tasks independently (i.e. we did not expect order effects). The complete procedure took approximately 35 min to complete.

2.2. Results and discussion

For all inferential analyses we used default Bayesian *t*-tests (Rouder et al., 2009) and Bayesian ANOVAs (Rouder et al., 2012). We did not have strong intuitions concerning effect sizes for the current work, so we relied on the default Cauchy priors in JASP (JASP Team, 2017).[4] A major advantage of Bayesian approaches over more conventional inferential analyses is that they allow for quantification of the statistical evidence in favor of *or* against the null hypothesis. This is an important

[4] Rouder et al. (2012) suggest that default Cauchy priors are appropriate for the analysis of many types of experimental designs. In order to check that our results do not depend on the selection of a specific prior, we examined the extent to which the Bayes factors obtained for comparisons of interest varied as a function of the prior when these were allowed to vary. The results, provided with our data, show that our key effects were robust across a range of Cauchy widths.

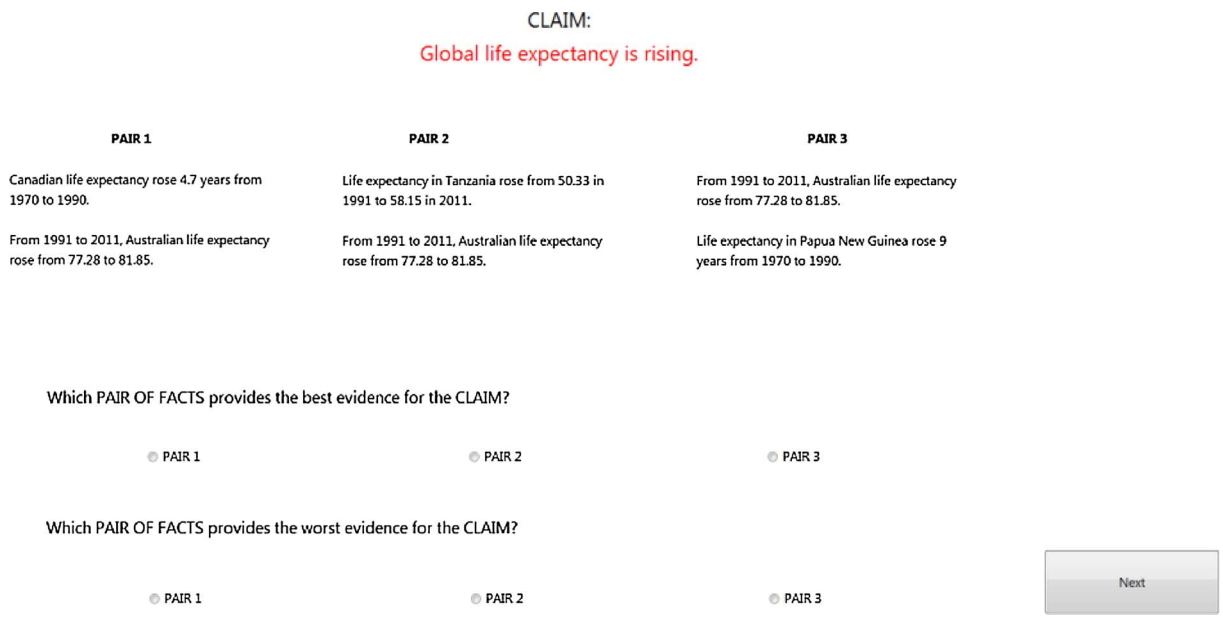

Fig. 1. Screenshot of the task for the claim "Global life expectancy is rising" showing evidence set 4 from Table 2.

advantage because it was of interest to identify both dimensions of diversity that increased or decreased perceived support for scientific arguments as well as dimensions that had little no effect on perceived support (i.e., the null hypothesis). The Bayes factor is directly interpretable as how many times more likely one hypothesis is to have generated the observed data over another. We use the notation BF_{10} to refer to Bayes factors where $BF_{10} > 1$ indicates support for the alternative hypothesis and $BF_{10} < 1$ support for the null hypothesis. For example, $BF_{10} = 10$ indicates the data is 10 times more likely to have come from the alternative hypothesis than the null hypothesis, and $BF_{10} = .1$ indicates the opposite conclusion. We follow the conventions suggested by Kass and Raftery (1995) that a BF between 3 and 20 (0.33–0.05) represents "positive" evidence for the alternative (or null) hypotheses respectively, a BF between 21 and 150 (0.049–0.0067) represents "strong" evidence and a BF above 150 (< .0067) represents "very strong" evidence.

2.2.1. Plausibility of scientific claims

We first examined the relative plausibility of the target claims, when they were evaluated in the absence of explicit evidence. Most conclusions were rated as likely to be true (mean plausibility > 75). The one exception was the claim "global rates of immunization of 1-year olds against Polio are rising", which was viewed as somewhat less plausible ($M = 69.17$, $SD = 21.71$) than other claims in the health domain. Hence overall, plausibility ratings were higher for the climate domain ($M = 83.83$, $SD = 14.65$) than the public health domain ($M = 78.52$, $SD = 13.68$). While the conclusion plausibility rating task and the diversity task were regarded as independent tasks, we also checked whether there was any carryover effect of the argument evaluation task on plausibility ratings. Plausibility ratings were somewhat higher when they were made after the argument evaluation task ($M = 82.5$, $SD = 12.39$) than when they were made before argument evaluations ($M = 79.90$, $SD = 16.05$). A 2 (task order) x (2) (conclusion domain) mixed effects Bayesian ANOVA was conducted to see whether ratings differed reliably across domains and whether completing the evidence diversity task before the conclusion plausibility rating task affected ratings. We found that the model including a main effect of conclusion domain performed the best relative to the null model, $BF = 3.92$ (1.05% error), though the evidence was relatively weak. This model was also more likely to have produced the data than a model also featuring an effect of order, $BF = 2.85$, but again the evidence was relatively weak. Completing the plausibility rating task after the evidence diversity task did not appear to have a noticeable effect on beliefs in the conclusions, though the evidence is somewhat ambiguous.

Given the positive evidence of a difference in conclusion plausibility across domains, we avoided direct comparison between domains in subsequent analyses of best-worst scores for diverse and non-diverse evidence. The main focus in these analyses was on the way that diverse evidence on each dimension affected acceptance of scientific claims within each domain.

2.2.2. Assessing the impact of diverse evidence using best minus worst scores (B-W)

A preliminary analysis found inconclusive to positive evidence that the pattern of best-worst choices was unaffected by whether these choices were completed before or after conclusion plausibility ratings (for the Climate context $BF_{10} < 1.15$ across dimensions; for the Public Health context, $BF_{10} < 0.35$ across dimensions). All subsequent analyses therefore collapsed across this factor.

The B-W scores for items containing each dimension of diversity in each domain are displayed in Fig. 2. Recall that positive numbers on this scale indicate a preference for claims based on diverse evidence, negative numbers indicate a preference for conclusions based on non-diverse evidence, and numbers close to 0 indicate that diversity did not influence acceptance of claims. The Figure shows that for the climate domain, geographically diverse but not socially diverse evidence was seen as increasing support for scientific claims. For the health domain, both geographically diverse and socially diverse evidence appeared to increase support for scientific claims. Somewhat surprisingly, it appeared that temporally diverse evidence had a negative effect on perceived support for scientific claims.

These effects were examined for each dimension using a series of Bayesian one sample *t*-tests. Each test compared the null hypothesis that the B-W scores for diverse evidence were equal to 0 (i.e. diversity had no influence on best and worst choices), to an alternative hypothesis that the B-W scores for diversity are different from zero, in either a positive or a negative direction. We used the default prior in JASP. In the climate domain, options featuring geographically diverse evidence were reliably chosen as the best type of support for scientific claims relating to climate change, as indicated by the positive B-W score in Fig. 2, $BF_{10} = 544.24$. In the climate domain, evidence diversity on the social dimension had little impact on perceived support for scientific

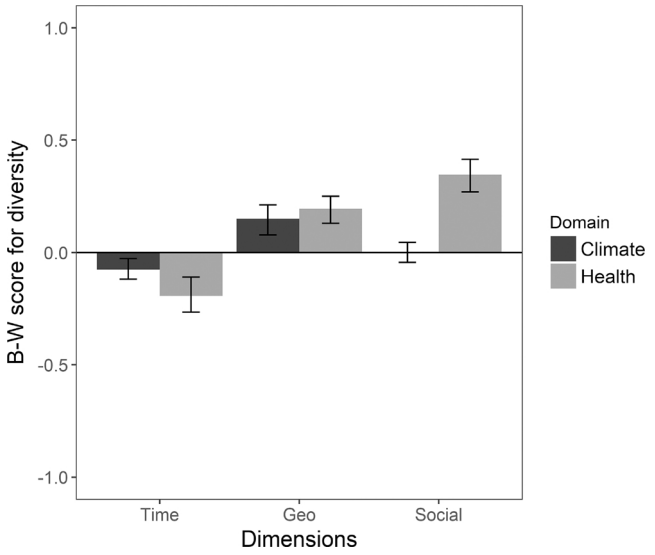

Fig. 2. Mean best-worst score (B-W) for the climate change and health arguments, collapsed across the 4 arguments in each condition. Scores above 0 indicate a preference for diverse sources of evidence whereas scores below 0 indicate a preference for non-diverse evidence (minimum score = −1, maximum score = +1). Error bars are 95% posterior credible intervals from the default prior (10 000 samples).

claims; i.e. there was no clear preference for choosing diverse arguments as the best evidential option, $BF_{10} = 0.14$. Surprisingly, on the temporal dimension, evidence diversity appeared to have a negative effect on belief in the scientific claim. In the climate domain, options containing temporally diverse evidence were most often selected as the worst form of evidence, $BF_{10} = 14.37$.

In the health domain we found that options containing diverse evidence on the geographical dimension, were more often chosen as the best form of evidence, $BF_{10} = 437497.2$. This was also true for options with diverse evidence on the social dimension, $BF_{10} = 1.03 \times 10^{11}$. Again, diversity on the temporal dimension had a negative effect on perceived support for scientific arguments strength with the diverse options often chosen as the worst form of evidence, $BF_{10} = 2667.478$.

We also examined the *relative* impact of diverse evidence on each of the three dimensions within each domain, by comparing the magnitudes of respective best-worst scores on different dimensions (e.g., for Fig. 2 climate data, we compared the magnitude of the three BW scores given in black). In the climate domain, options with geographically diverse evidence were chosen as providing the best support for scientific conclusions more often than options with socially diverse evidence, $BF_{10} = 387.42$. Within this domain there was inconclusive evidence for a difference in the choices of options with socially diverse evidence and those with temporally diverse evidence, $BF_{10} = 1.46$.

A different pattern was found in the health domain where options with socially diverse evidence were chosen as the best form of support more often than those with geographically diverse evidence, $BF_{10} = 29.3$. Likewise, socially diverse evidence had more of an impact on perceived support for scientific arguments than temporally diverse evidence, $BF_{10} = 2.49 \times 10^{12}$. The same was true of geographical evidence, $BF_{10} = 1.08 \times 10^{9}$.

These results show that evidential diversity on the target dimensions did affect lay people's beliefs in scientific claims. Notably however, the types of diverse evidence that led to increased belief in these claims differed between domains. Evidence that was geographically diverse was seen as providing the strongest support for conclusions relating to global climate change. The socio-cultural diversity of evidence had little effect on beliefs in this domain. In the public health domain, although both social and geographical diversity of scientific evidence had a positive impact on support for scientific arguments, the strongest form of support was provided by socially diverse evidence.

Consistent with previous laboratory investigations of inductive reasoning, these data show that evidential diversity is a potent factor in lay perceptions of the validity of scientific claims. The important novel finding, however, is that in the more complex domain of factual scientific data this effect is selective and varies across domains.

A somewhat surprising finding was that in each domain *less* diverse evidence on the temporal dimension was seen as providing greater support for scientific conclusions than evidence drawn from diverse time-periods. There are at least two possible explanations for this finding. The first is that people may be skeptical about accepting data drawn from very different time-periods. Such data are more likely to be based on the application of different scientific methods, which could reasonably be expected to change over time, and this may undermine the perceived continuity of the data. A related idea is that people may believe that the relevant causal factors (e.g., large increases in industry based carbon emissions) are restricted to specific time periods (e.g., the past 50 years) and so that data repeatedly sampled from this period are more relevant than data from more widely dispersed periods. Such reasoning about specific causal connections between argument premises and conclusions can override the heuristic effect of evidential diversity in studies of inductive reasoning (e.g., Feeney and Heit, 2011; Medin et al., 2003).

Alternatively, this effect may reflect a more general preference for basing conclusions on more recent data. Given that argument conclusions in both climate and health domains were about expected future trends, many people may have believed that recent data was most relevant to assessing the strength of these conclusions.

It was not possible to separate out these two explanations in the current experiment. Here more diverse arguments always involved a combination of data from a relatively recent period with "older" data (see supplementary materials), whereas less diverse arguments involved the combination of two statements describing relatively recent data (on the majority of items) or two statements describing more temporally distant data (on a minority of items). A further complication was that the duration of the time samples was not standardized across items. These issues were addressed in Experiment 2 with the aim of differentiating between explanations of the temporal non-diversity result.

3. Experiment 2

In Experiment 2 non-diverse evidence on the temporal dimension came in two forms; "recent", where arguments contained data from recent time-periods (1990–2010), and "past", where arguments contained data from more remote periods (1960–1980). As before, diverse temporal evidence always combined past and recent data. All temporal data were of the same duration (10 years). If preference for non-diverse temporal evidence found in Experiment 1 was due to skepticism about the continuity of scientific data sampled from diverse time-periods, then we should see the same preference for arguments with non-diverse premises in both the past and recent versions. If however, it was based on a preference for conclusions based on more recent data, then the temporal non-diversity effect should only appear in the recent data version. Although there were some other minor adjustments to arguments (see below) the way that evidence diversity was manipulated on the geographical and social dimensions remained largely unchanged from the earlier experiment.

A further novel feature was the inclusion of a questionnaire about political affiliations and beliefs about climate change. This was added because there is evidence that acceptance of climate change science is mediated by left-right political orientation (see Newell et al., 2014 for a review). In some cases the effect of political orientation has been argued to have more impact on climate change beliefs (and many other scientific issues such as evolution) than comprehension of the relevant science (Kahan, 2015; Kahan et al., 2012). In the climate domain therefore, those with a more right-wing

political orientation were expected to be generally less likely to endorse claims about global warming. It is unclear, however, whether the effects of political orientation will extend to the way that people use diverse evidence to evaluate a scientific claim. Conventionally, evidence diversity is seen as a general reasoning heuristic that can be used to evaluate the strength of an inductive argument independent of its specific content (Hayes and Heit, 2013). Whether or not use of this heuristic in the climate domain is affected by political orientation remains to be seen.

3.1. Method

3.1.1. Participants

179 Australian first year Psychology students participated for course credit. The mean age was 19.23 (SD = 3.57) and there were 130 females. The experiment featured 2 between-subjects conditions (defined below) − past evidence (n = 92) and recent evidence (n = 87).

3.1.2. Design and materials

As in Experiment 1, all participants made choices about evidence options that varied in geographical and social diversity as shown in Tables 1 and 2. The main change to the design was in the way that temporal diversity was manipulated. In this study, participants were randomly assigned to one of two groups in which they received items with the fixed premise either containing past (years 1960–1980) or recent (years 1990–2010) data on the temporal dimension. Hence, for a given participant, non-diverse evidence took the form of two samples from either the past (e.g., In Australia, each decade from 1960 to 1980 was warmer than the one before + During 1960 to 1980, New Zealand temperatures increased by 0.6 °C) or from more recent periods (e.g., In Australia, during 1990 to 2010, the average temperature has been 0.5 °C warmer than the global long term average + The mean temperature of New Zealand during 1990–2010 was 0.26 °C above the historical average). Temporally diverse arguments included evidence from each period (e.g., In Australia, each decade from 1960 to 1980 was warmer than the one before + The mean temperature of New Zealand during 1990–2010 was 0.26 °C above the historical average). The second premises of all arguments were the same across the manipulation, but presented in different combinations with the fixed premise to satisfy the option structures presented in Table 2. For example, using the alternative fixed premises relating to temperature increases above, the premise "Temperatures in Indonesia rose 0.14 °C from 1990 to 2010" was paired with the fixed premise to form option [0 0 1] in the recent condition and option [1 0 1] in the past condition, while the premise "Average maximum temperatures in Papua New Guinea increased by 0.11 °C per decade from 1960 to 1980" was paired with the fixed premise to create option [1 0 1] in the recent condition and [0 0 1] in the past condition. In effect, this meant that all participants received the same information about the geographical and social dimensions. Hence, the between-subjects manipulation was only relevant to the temporal dimension.

As in Experiment 1, we based all premises on scientific facts taken from authoritative sources. In addition to the changes in the time dimension, minor adjustments were made to reduce the size of numerical differences between statements presented in alternative options. The full list of arguments can be found in the supplementary materials. To check that participants were engaged in the task, we also included items with claims known to be scientifically false (e.g., "The volume of glaciers across the globe is increasing") in the conclusion plausibility ratings part of the experiment.

There were also several measures of political affiliation and attitudes towards climate change. For our measure of political affiliation, we asked participants how likely they were to ever vote (on a scale from 1 (Very Unlikely) to 10 (Very Likely)) for each of three Australian Federal political parties in a general election: The center right Liberal Party, the center left Labor Party, and the environmentalist Greens party. We also asked participants to choose which of four alternatives best described their beliefs about climate change: a) they believed that anthropogenic global warming was occurring, b) they believed in global warming but that the cause was not anthropogenic, c) they did not believe global warming was occurring, d) they did not know whether warming was occurring. In this section, we also asked participants whether they were Australian citizens, but this data was not considered in our analysis.

3.1.3. Procedure

The total number of items completed by individual participants was the same as in Experiment 1. Unlike Experiment 1, conclusion plausibility ratings were always completed prior to best-worst assessment of arguments, as we did not find evidence that the order manipulation made a difference to the B-W scores in Experiment 1. Political and climate change attitude measures were collected after argument assessments.

3.2. Results and discussion

3.2.1. Plausibility of scientific claims

As in Experiment 1, the scientific claims used in the study were generally rated as highly likely, except for the health argument concerning increases in immunization rates (M = 63.65, SD = 22.09). Climate change conclusions (M = 86.19, SD = 13.45) were rated as more plausible than health conclusions (M = 72.33, SD = 12.06), although the effect was stronger than in Experiment 1, $BF_{10} = 4.73 \times 10^{24}$. Participants rated test items as more plausible than the scientifically false items in both the climate (M_D = 52.85, SD_D = 23.21), $BF_{10} > 10^{38}$, and public health domains, (M_D = 32.8, SD_D = 23.92), $BF_{10} > 10^{38}$.

3.2.2. Assessing the impact of diverse evidence using best worst scores (B-W scores)

Fig. 3 shows mean B-W scores for climate and health options that varied in evidential diversity on the three key dimensions. For inferences concerning geographical and social diversity, we collapsed

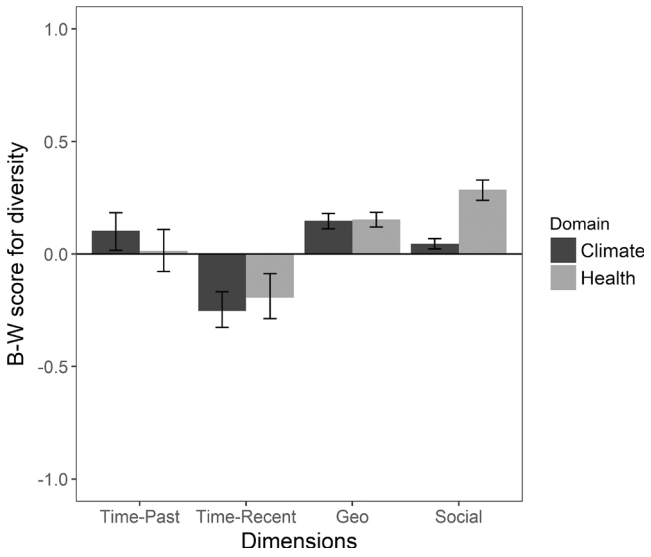

Fig. 3. Mean best-worst score (B-W) for the climate change and health arguments, collapsed across the 4 arguments in each condition. Scores above 0 indicate a preference for diverse sources of evidence whereas scores below 0 indicate a preference for non-diverse evidence. Arguments on the social and geographical dimensions include data from all participants (N = 179); arguments on the temporal dimension are separated into those that used past (1960–1980) (n = 92) and recent (1990–2010) (n = 87) non-diverse data. Error bars are 95% posterior credible intervals from the default prior (10 000 samples).

across the between-subjects fixed premise manipulation since we had no expectation that the manipulation of temporal information would influence preference for diversity of the other dimensions ($N = 179$ for these comparisons). We again first analyzed these data by examining whether arguments with diverse evidence tended to be selected as the best or worst form of evidence. Consistent with Experiment 1, options featuring geographically diverse evidence in the climate domain were reliably chosen as the best form of evidence, $BF_{10} = 4.18 \times 10^{11}$. In this case however, options with socially diverse evidence were also chosen as the best form of support for climate claims, $BF_{10} = 116.02$. There was negligible impact of temporal diversity on B-W scores when the non-diverse options were based on "past" evidence ($n = 92$), $BF_{10} = 1.81$. However, those completing the "recent" evidence version generally chose diverse evidence as the worst form of support for climate claims ($n = 87$), $BF_{10} = 1.7 \times 10^6$. In other words, two statements describing recent non-diverse data were generally seen as more compelling support for scientific claims about climate change than the more diverse mixture of past and recent data.

We also replicated the main results from Experiment 1 regarding diverse evidence in the public health domain, with strong evidence that socially diverse premises were selected as the best form of evidence, $BF_{10} = 2.15 \times 10^{23}$. The same was true of geographically diverse premises, $BF_{10} = 5.52 \times 10^{13}$. We found evidence in favor of the null hypothesis with regard to the B-W score for temporal diversity in the past evidence condition ($n = 92$), $BF_{10} = 0.12$. In contrast, and like the climate condition, temporally diverse premises were generally selected as the worst form of evidence in the recent evidence condition ($n = 87$), $BF_{10} = 62.9$.

We again examined the relative impact of evidence diversity by comparing the magnitude of best-worst scores between dimensions within each domain. In the climate domain, geographically diverse evidence was seen as providing stronger support for scientific claims (i.e. had higher B-W scores) than socially diverse evidence, $BF_{10} = 19053.77$. In the public health condition socially diverse evidence was seen as providing stronger support for claims than arguments based on geographical diversity, $BF_{10} = 75747.72$. This replicates the key trends for these dimensions found in the earlier study.

For comparisons of preference between the temporal and other dimensions, we performed our analysis within each condition of the between-subjects past vs. recent evidence manipulation. In the climate domain, in the past-data condition ($n = 92$), we found evidence in favor of the null hypothesis that there was no difference in B-W scores between the temporal and geographical dimensions, $BF_{10} = 0.19$, nor between the temporal and social dimensions, $BF_{10} = .17$. In contrast, in the recent-data condition ($n = 87$) we found strong evidence in favor of a difference between temporal and geographical B-W scores, $BF_{10} = 3.3 \times 10^{10}$, and temporal and social B-W scores, $BF_{10} = 1.4 \times 10^6$. Temporal diversity in recent data had a negative effect on perceived support for scientific claims, whereas diversity on the other two dimensions had a positive effect (see Fig. 3).

In the health domain when the arguments contained past data ($n = 92$), geographical diversity was seen as a stronger form of evidence than temporal diversity, $BF_{10} = 3.14$. Likewise, social diversity was seen as a stronger form of evidence than temporal diversity, $BF_{10} = 329.90$. In arguments containing recent data ($n = 87$), we found that B-W scores for both geographical, $BF_{10} = 1087411$, and social diversity, $BF_{10} = 5.03 \times 10^{11}$, were different from and opposite in sign to scores for temporal diversity. Fig. 3 shows that in both of these cases, temporally diverse evidence was generally chosen as the worst form of evidence (with a negative B-W score in Fig. 3) whereas diversity on the geographical and social dimensions had a positive effect on belief in scientific claims.

In the main, the effects of the evidence diversity manipulation on the strength of scientific arguments for the geographical and social dimensions were similar to those found in Experiment 1. Evidence options including diverse geographical evidence were seen as providing the strongest support for scientific claims in the climate domains. In this case, social diversity also had a small but reliable positive effect on the perceived support for claims about climate change. This may have occurred because the sample size for geographical and socially diverse dimensions was essentially double the size from Experiment 1. Hence, with a larger sample size we found evidence for an influence of social diversity on preference for climate claims that we did not detect in Experiment 1. Nevertheless, in this domain, geographically diverse evidence was still viewed as providing stronger evidence for scientific claims than socially diverse evidence.

Once again, the results concerning evidence diversity on the temporal dimension differed from those on the other two dimensions. People saw temporal evidence drawn from recent periods (i.e. recent non-diverse evidence) as more persuasive support for climate and public health conclusions than diverse temporal evidence. In contrast, including older data had little impact. This suggests that the Experiment 1 finding of a preference for arguments based on non-diverse temporal data was largely driven by a preference for recent data over older data (and over mixtures of old and recent data).

3.2.3. Political and climate change attitudes

89.83% of the sample believed that climate change was real and caused by human actions. The sample was "left leaning" with participants most willing to vote for the center-left Labor Party ($M = 5.95$ out of 10, $SD = 2.58$), followed by the environmentalist The Greens ($M = 4.84$, $SD = 3.24$) and then the center-right Liberal Party ($M = 4.34$, $SD = 2.96$). We ran a Bayesian linear regression (cf. Rouder and Morey, 2012) with default priors on the covariates to see if conclusion plausibility ratings for climate items were influenced by climate change beliefs and political affiliation. Bayesian linear regression takes a model comparison approach, where we obtain the likelihood of the data under each possible model (i.e. the model containing all predictors and interactions, or only some predictors, or the intercept only model) and compare the likelihoods using Bayes factors. The response options on the climate change question were coded categorically as a predictor variable. A regression model containing only climate change beliefs as a predictor of conclusion plausibility had the highest Bayes Factor relative to the null model, $BF_{10} = 22.75$. There is inconclusive evidence that the climate change belief model performed better than a model which also included willingness to vote for the center-right party as an additional predictor, $BF = 1.39$. The model that only contained climate change beliefs outperformed the remaining models by a Bayes factor of at least $BF = 5.08$ (e.g. the model containing only climate change belief accounted for the data better than a model containing only willingness to vote for the center-right party, $BF = 11.06$). Those with strong a-priori beliefs in anthropogenic climate change were more likely to endorse scientific claims about current and future global warming, but there was no clear evidence that political orientation predicted acceptance of such claims when included in a model with climate change beliefs.[5]

To see whether the climate attitudes and political measures affected the way that people used diverse evidence in climate items, we also ran separate regression analyses with B-W scores for each dimension (time-past, time-recent, geographical and social) as the dependent measures. In the climate context, the best performing model against the null across all dimensions was the climate change beliefs model, but the evidence was either inconclusive or there was

[5] For completeness, we also ran regression analyses examining prediction of plausibility ratings and B-W scores in health items from climate attitude and political affiliation. There was equivocal evidence that Climate change beliefs predicted endorsement of public health conclusions, $BF_{10} = 1.30$, and that willingness to vote for the center-left party predicted B-W scores for geographical diversity, $BF_{10} = 1.22$. All other models have BF_{10}'s < 0.65 compared to the null model.

positive evidence in favor of the null (BF_{10} between 0.33 and 0.69). Although environmental beliefs predicted overall endorsement of conclusions about global warming, they did not appear to predict whether participants were influenced by evidential diversity in climate items. However, more data (and/or a more politically diverse sample) would be required to be confident in this conclusion, as well as the role of political orientation in accepting scientific claims about global warming.

4. General discussion

Two studies examined the impact of evidential diversity on perceived support for scientific claims. The structure of the scientific arguments used in this experiment was more complex than those used in previous laboratory studies of inductive inference, in that they contained three different types of evidence (temporal, geographical, social), each of which could be presented as diverse or non-diverse.

As in previous laboratory work with fictional categorical materials (Heit et al., 2004), we found that diverse evidence was often seen as a stronger basis for drawing scientific conclusions than non-diverse evidence. An important novel finding however, was that the impact of diversity was selective. The type of diverse evidence that had the greatest impact differed between domains. In the climate domain, geographically diverse evidence had the largest effect in perceived support for scientific conclusions. Socially diverse evidence had a smaller positive effect on perceived support for climate conclusions. The reverse pattern was found in the health domain where socially diverse evidence had the strongest positive impact on belief in the scientific conclusion.

The selective nature of the diversity effects that we observed are most likely due to differences in lay mental models of the relevant causal factors in each domain (Bostrom et al., 2012; Newell et al., 2014). In the case of climate change, evidence of the climatic effects of CO_2 accumulation from diverse geographical sites could be seen as supporting the conclusion that warming is indeed a global rather than a local phenomenon. The causal connection between social diversity and climate change is less clear, and hence evidence diversity on this dimension had less of an impact on belief in relevant scientific conclusions. In contrast, lay theories of health and illness (Hagger and Orbell, 2003; Turk et al., 1986) typically emphasize interactions between individual physiology and local social and environmental factors such as levels of nutrition and hygiene. In this context, the observation that a public health intervention such as immunization has a positive effect across diverse socio-cultural environments represents strong evidence of its general efficacy.

Public acceptance of scientific data on climate change is often influenced by one's political orientation and general environmental attitudes (e.g., Kahan, 2015; Newell et al., 2014). Consistent with this work we found that prior belief in anthropogenic climate change was predictive of endorsement of the scientific claims concerning climate change presented during the experiment. Notably however, the tendency to use geographical diversity as a way of evaluating the strength of claims about climate change was unaffected by such prior beliefs. We need to be cautious about this finding since our undergraduate sample is likely to be biased in favor of climate change science relative to the general population. Nevertheless, it suggests that the use of evidence diversity as a heuristic for evaluating the scientific evidence operates regardless of one's prior knowledge of or general attitudes to a controversial scientific topic such as climate change.

Naturally, the public discourse on controversial topics such as climate change often involves exposure to information that distorts or contradicts consensual scientific views (Oreskes and Conway, 2010). An important challenge for future work is to examine whether strengthening belief in scientific data through manipulations like evidence diversity can increase *resistance* to such misinformation (cf. van der Linden et al., 2017).

An unexpected finding from both studies was that evidence diversity on the temporal dimension had a *negative* influence on belief in scientific claims in both climate and health domains. Experiment 2 showed that this was due to a preference for more recent data over a combination of more and less recent data. This result may reflect the application of a heuristic rule that recent data is more relevant to conclusions about current/future states than older evidence (or mixtures of old and new evidence). Such a heuristic however, would be inconsistent with the reasoning of many scientific experts. In the climate domain for example, observation of increased temperatures and rising sea levels over a long time scale is seen as strengthening the conclusion that the planet is warming and that such warming is likely to continue (e.g., Crowley, 2000; IPCC, 2014). Our reasoning results concerning the temporal dimension are broadly consistent with work which shows that people often have difficulty representing and reasoning about phenomena which occur over long time periods (e.g., Resnick et al., 2012) or which involve change in quantitative variables over time (e.g., Newell et al., 2016; Sterman and Booth Sweeney, 2007).

A further reason why our manipulation of temporal diversity may have not been persuasive is that we presented evidence on absolute changes in temperature, sea level and so on, within particular time-periods. In domains like climate change however, the most compelling forms of temporal evidence are likely to involve comparing *rates of change* over time (e.g., Mann et al.'s, 1998, "hockey stick" data on long-term trends in surface temperature). In this context, it is the recent rate of climatic change that is relevant for predicting future trends.

4.1. Implications for science communication

The current results have some important implications for communicating scientific results to members of the public. If the goal is to convey the strength of evidence for a particular scientific conclusion (e.g., that the global climate is warming) then presenting diverse forms of evidence is likely to be effective, as long as that evidence is also seen as having a causal connection with the target phenomenon. Concretely, in the case of climate change, science communicators would be well advised to emphasize the convergent trends in data on anthropogenic global warming obtained from both local *and* geographically disparate locations, rather than just highlighting either global or local trends (cf. Painter, 2010). In the case of public health, emphasizing the convergence between data obtained from local locations and those that are geographically and culturally dissimilar may lead to greater acceptance of the relevant science.

More broadly, we see the current results as highlighting the relevance of basic research on inductive reasoning for the framing of scientific communication. Such research suggests factors that can be manipulated when presenting scientific data to strengthen (or weaken) their acceptance by a lay audience. The current studies focused on just one of these factors – diversity. The induction literature offers a potentially rich source of additional factors for future study.

Appendix A

See Table A1.

Table A1
Target conclusions for each domain. Note that the conclusion in Experiment 1 "The incidence of diabetes is increasing across the globe" was replaced by "Average Body Mass Index (BMI) is increasing across the globe" in Experiment 2. All other conclusions remained the same across both experiments.

Domain		
Conclusion #	Climate	Health
1	The global climate is warming	Experiment 1: The incidence of diabetes is increasing across the globe Experiment 2: Average Body Mass Index (BMI) is increasing across the globe
2	Human activity is contributing to the accumulation of CO2 in the atmosphere	Global life expectancy is rising
3	Global sea levels are rising	The use of modern contraceptives among married women (aged 15–49) is rising across the globe
4	Extreme weather events are increasing in frequency across the globe	Global rates of immunization of 1-year olds against Polio are rising.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.gloenvcha.2018.01.004.

References

Bacon, F., 1620/1898. Novum Organum. George Bell & Sons, London.

Bostrom, A., O'Connor, R.E., Böhm, G., Hanss, D., Bodi, O., et al., 2012. Causal thinking and support for climate change policies: international survey findings. Glob. Environ. Change 22, 210–222.

Brulle, R.J., Carmichael, J., Jenkins, J.C., 2012. Shifting public opinion on climate change: an empirical assessment of factors influencing concern over climate change in the US, 2002–2010. Clim. Change 114, 169–188.

Carey, S., 1995. On the origins of causal understanding. In: Sperber, D., Premack, D., Premack, A.J. (Eds.), Causal Cognition: A Multi-Disciplinary Approach. Clarendon Press, Oxford, UK, pp. 268–308.

Cohen, E., 2009. Applying best-worst scaling to wine marketing. Int. J. Wine Bus. Res. 21, 8–23.

Corner, A., Hahn, U., 2009. Evaluating scientific arguments: evidence, uncertainty & argument strength. J. Exp. Psychol.: Appl. 15, 199–212.

Crowley, T.J., 2000. Causes of climate change over the past 1000 years. Science 289 (5477), 270–277.

Feeney, A., Heit, E., 2011. Properties of the diversity effect in category-based inductive reasoning. Think. Reason. 17, 156–181.

Freudenburg, W.R., Muselli, V., 2010. Global warming estimates, media expectations: and the asymmetry of scientific challenge. Glob. Environ. Change 20, 483–491.

Fujita, K., Clark, S.L., Freitas, A.L., 2014. Think globally, act locally: construal levels and environmentally relevant decision-making. In: van Trijp, H.C.M. (Ed.), Encouraging Sustainable Behavior: Psychology and the Environment. Psychology Press, New York, pp. 87–107.

Hagger, M.S., Orbell, S., 2003. A meta-analytic review of the common-sense model of illness representations. Psychol. Health 18, 141–184.

Hahn, U., Bailey, T.M., Elvin, L.B., 2005. Effects of category diversity on learning, memory, and generalization. Mem. Cognit. 33, 289–302.

Hahn, U., Harris, A.J.L., Corner, A., 2016. Public reception of climate science: coherence: reliability and independence. Top. Cognit. Sci. 8, 180–195.

Hayes, B.K., Heit, E., 2013. Induction. In: Reisberg, D. (Ed.), Oxford Handbook of Cognitive Psychology. Oxford University Press, New York, USA, pp. 618–634.

Hayes, B.K., Goodhew, A., Heit, E., Gillan, J., 2003. The role of diverse instruction in conceptual change. J. Exp. Child Psychol. 86, 253–276.

Hayes, B.K., Heit, E., Swendsen, H., 2010. Inductive reasoning. Wiley Interdiscip. Rev. Cognit. Sci. 1, 278–292.

Heit, E., Hahn, U., 2001. Diversity-based reasoning in children. Cognit. Psychol. 43, 243–273.

Heit, E., Hahn, U., Feeney, A., 2004. Defending diversity. In: Ahn, W., Goldstone, R.L., Love, B.C., Markman, A.B., Wolff, P. (Eds.), Categorization Inside and Outside of the Lab: Festschrift in Honor of Douglas L. Medin. American Psychological Association, Washington, DC, pp. 2004.

Hempel, C.G., 1966. Philosophy of Natural Science. Prentice-Hall, Englewood Cliffs, N.J.

IPCC, 2014. In: Core Writing Team, Pachauri, R.K., Meyer, L.A. (Eds.), Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva, Switzerland.

JASP Team, 2017. JASP (Version 0.8.4) [Computer Software].

Jones, L.G., Hawkins, G.E., Brown, S.D., 2015. Using best-worst scaling to improve psychological service delivery: an innovative tool for psychologists in organized care settings. Psychol. Serv. 12, 20–27.

Kahan, D.M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L.L., Braman, D., Mandel, G., 2012. The polarizing impact of science literacy and numeracy on perceived climate change risks. Nat. Clim. Change 2, 732–735.

Kahan, D.M., 2015. Climate-science communication and the measurement problem. Polit. Psychol. 36 (S1), 1–43.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.

Keil, F.C., 2003. Folkscience: coarse interpretations of a complex reality. Trends Cognit. Sci. 7, 368–373.

Kim, N.S., Keil, F.C., 2003. From symptoms to causes: diversity effects in diagnostic reasoning. Mem. Cognit. 31, 155–165.

López, A., 1995. The diversity principle in the testing of arguments. Mem. Cognit. 23, 374–382.

Leviston, Z., Greenhill, M., Walker, I.A., 2015. Australian Attitudes to Climate Change and Adaptation: 2010–2014. CSIRO, Perth, Australia.

Lo, Y., Sides, A., Rozelle, J., Osherson, D., 2002. Evidential diversity and premise probability in young children's inductive judgment. Cognit. Sci. 26, 181–206.

Louviere, J.J., Street, D., Burgess, L., Wasi, N., Islam, T., Marley, A.A., 2008. Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information. J. Choice Modell. 1, 128–164. http://dx.doi.org/10.1016/S1755-5345(13)70025-3.

Louviere, J.J., Flynn, T.N., Marley, A.A.J., 2015. Best-Worst Scaling: Theory, Methods and Applications. Cambridge University Press.

Mann, M.E., Bradley, R.S., Hughes, M.K., 1998. Global-scale temperature patterns and climate forcing over the past six centuries. Nature 392, 779–787.

Marley, A.A.J., Islam, T., Hawkins, G.E., 2016. A formal and empirical comparison of two score measures for best-worst scaling. J. Choice Modell. 21, 15–24. http://dx.doi.org/10.1016/j.jocm.2016.03.002.

McDonald, R., Chai, H.-Y., Newell, B.R., 2015. Personal experience and the 'psychological distance' of climate change: an integrative review. J. Environ. Psychol. 44, 109–118.

Medin, D.L., Coley, J., Storms, G., Hayes, B.K., 2003. A relevance theory of induction. Psychonomic Bull. Rev. 10, 517–532.

Newell, B.R., McDonald, R.I., Brewer, M., Hayes, B.K., 2014. The psychology of environmental decisions. Ann. Rev. Environ. Res. 39, 443–467.

Newell, B.R., Kary, A., Moore, C., Gonzalez, C., 2016. Managing the budget: stock-flow reasoning and the CO2 accumulation problem. Top. Cognit. Sci. 8, 138–159.

Oreskes, N., Conway, E.M., 2010. Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming. Bloomsbury, New York, NY.

Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A., Shafir, E., 1990. Category-based induction. Psychol. Rev. 97, 185.

Painter, J., 2010. Summoned by Science: Reporting Climate Change at Copenhagen and Beyond. Reuters Institute for the Study of Journalism, Oxford, UK.

Ranney, M.A., Clark, D., 2016. Climate change conceptual change: scientific information can transform attitudes. Top. Cognit. Sci. 8, 49–75.

Resnick, I., Atit, K., Shipley, T.F., 2012. Teaching geologic events to understand geologic time. In: Kastens, K.A., Manduca, C.A. (Eds.), Earth and Mind II: A Synthesis of Research on Thinking and Learning in the Geosciences: Special Papers 486. Geological Society of America, pp. 41–43.

Rhodes, M., Gelman, S.A., Brickman, D., 2010. Children's attention to sample composition in learning, teaching and discovery. Dev. Sci. 13 (3), 421–429.

Rouder, J.N., Morey, R.D., 2012. Default Bayes factors for model selection in regression. Multivar. Behav. Res. 47, 877–903.

Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G., 2009. Bayesian *t* tests for accepting and rejecting the null hypothesis. Psychonomic Bull. Rev. 16 (2), 225–237.

Rouder, J.N., Morey, R.D., Speckman, P.L., Province, J.M., 2012. Default bayes factors for ANOVA designs. J. Math. Psychol. 56, 356–374.

Salmon, W.C., 1984. Scientific Explanation and the Causal Structure of the World. Princeton University Press, Princeton, N J.

Shafto, P., Coley, J.D., Vitkin, A., 2007. Availability in category-based induction. In: Feeney, A., Heit, E. (Eds.), Inductive Reasoning: Experimental, Developmental, and Computational Approaches. Cambridge University Press, New York, NY, pp. 114–136.

Sinatra, G.M., Hofer, B.K., 2016. Public understanding of science: policy and educational implications. Policy Insights Behav. Brain Sci. 3, 245–253.

Sloman, S., Lombrozo, T., Malt, B., 2007. Ontological commitments and domain specific categorization. In: Roberts, M.J. (Ed.), Integrating the Mind. Psychology Press, New

York, NY, pp. 105–220.
Smith, N., Leiserowitz, A., 2012. The rise of global warming skepticism: exploring affective image associations in the United States over time. Risk Anal. 32, 1021–1032.
Sterman, J.D., Booth Sweeney, L., 2007. Understanding public complacency about climate change: adults' mental models of climate change violate conservation of matter. Clim. Change 80, 213–238.
Street, D.J., Burgess, L., Louviere, J.J., 2005. Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. Int. J. Res. Mark. 22, 459–470.
Sutton, R., Suckling, E., Hawkins, E., 2015. What does global mean temperature tell us about local climate? Philos. Trans. R. Soc. A 373, 20140426.
Turk, D.C., Rudy, T.E., Salovey, P., 1986. Implicit models of illness. J. Behav. Med. 9, 453–474.
van der Linden, S., Leiserowitz, A., Rosenthal, S., Maibach, E., 2017. Inoculating the public against misinformation about climate change. Glob. Challenges 1, 1–7.
Wayne, A., 1995. Bayesianism and diverse evidence. Philos. Sci. 62, 111–121.
Weber, E.U., 2010. What shapes perceptions of climate change? Wiley Interdiscip. Rev. Clim. Change 1, 332–342.